

Dell APEX Cloud Platform for Red Hat OpenShift: An easily deployable and powerful solution to jumpstart your next AI innovation

The 4th Generation Intel Xeon Scalable processor-powered solution deployed in less than two hours and ran a Kubernetes container-based generative AI workload effectively

With official Red Hat and Dell documentation as guides, we easily deployed the necessary cloud infrastructure to run a Llama2, a large language model (LLM).

LLMs parse and generate human-like text, which many organizations could use for multiple practical applications:



Retail
Help customers with better support chatbots



Marketing
Speed content, ideas, and edits



Manufacturing
Analyze customer feedback for product design and manufacturing improvements



Healthcare
Enhance clinical decisions



Cybersecurity
Map regulations to policies and controls



Dell APEX Cloud Platform + 4th Generation Intel Xeon Scalable processors + Red Hat OpenShift AI

- Powerful resources and infrastructure
- Single pane of glass management with OpenShift Web Console



Llama 2 + Redis + Gradio

- Llama 2 is a pre-trained LLM
- Redis served as a document index
- Gradio was the graphical interface for user



Functional LLM that answers queries near instantly

- Easy to deploy
- Less than 2 hours to useable GenAI output

To learn more, [read the report](#) and the [science behind the report](#)